REMARKS

The Advisory Action mailed on March 1, 2007 has been given careful consideration by applicant. Reconsideration of the application is requested in view of the amendments and comments herein. Claims 1, 3, 10-12, 16, 18, and 19-21 have been amended. Claims 2 and 5 have been cancelled and claim 22 has been added.

The Office Action

Claims 1, 10-13 and 15 are rejected under 35 U.S.C. §102(b) as being anticipated by Kubota (US Patent No. 6,041,323);

Claims 3-7 and 16-21 are rejected under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Gilfillan et al. (US PG Pub. No. 2002/0165856);

Claims 8-9 are rejected under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Withgott et al. (US Patent No. 5,748,805); and

Claim 14 is rejected under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Cofino et al. (US PG Pub. No. 2005/0187931).

Anticipation Rejection

The Examiner has rejected claims 1-2, 10-13 and 15 under 35 U.S.C. §102(b) as being anticipated by Kubota (US Patent No. 6,041,323). This rejection should be withdrawn for at least the following reasons. Kubota does not teach or suggest the subject invention as set forth in the subject claims.

Independent claim 1 recites a method for identifying output documents similar to an input document. Lists of keywords are identified for each output document in the first set of documents by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages. A measure of similarity is computed between the input document and each output document in the first set of documents. A second set of documents is defined with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value. The list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document. The query is repeated until a predetermined number of results are obtained or the query is terminated. If the second set of documents includes a matching document but no similar documents, the above steps are repeated using the matching document to identify similar documents. Kubota does not teach or suggest such claimed aspects of the subject invention.

More particularly, Kubota does not teach or suggest tokenizing one or more keywords employed in a search at one or more predefined word boundaries while maintaining order of the sequence of input text, as recited in the subject claims. As known, tokenization is the process of taking categorized blocks of text, usually consisting of indivisible characters known as lexemes that are categorized according to function, giving them meaning. Kubota does not teach such tokenization. Instead Kubota teaches extracting unique character strings from an input document and a performing a similarity search by using the unique character string. The extraction of the unique character string is performed by calculating and evaluating an amount of feature of a character string through comparison between appearance frequency appearing in the input document and appearance frequency in a set of documents to be searched. See, e.g., Abstract. There is no mention of taking categorized blocks of text according to function to give them meaning. Accordingly, Kubota does not contemplate tokenization of any keywords to perform this search.

Moreover, Kubota does not teach or suggest the translation of keywords into one or more languages, as recited in the subject claims. Rather, Kubota performs searches based on character strings from documents. Such documents are presented in a single language and Kubota is silent regarding expanding a search to include translations of the selected character strings.

Additionally, Kubota does not teach or suggest repeating a query until a predetermined number of results are obtained or the query is terminated. Instead, Kubota teaches determining whether all fixed length chains created from the search character string have been searched. If so, the process proceeds to a next step. If not, the process returns to a previous step where the search process is performed for the next fixed length chain by using the character chain file. Thus, Kubota teaches performing searches until all search terms have been searched. There is no mechanism disclosed that contemplates performing such a search until a predetermined number of results have been located or until a search is terminated.

In addition, Kubota does not mention if the second set of documents includes a matching document but no similar documents repeating a search using the matching document to identify similar documents. The general measure of quality of search results has two components: precision and recall. Search results have good precision when the document being searched for is identified. Search results have good recall when not only the document being search for is identified but also all copies of the document being search for are identified. In this manner, in evaluating search results for input documents that were introduced via OCR or known to be a partial document, search results tend to have good precision but poor recall. When this occurs, as recited in the subject claims, the search is repeated using the matching document to identify similar documents when the second set of documents includes a matching document but no similar documents. The search is performed using the identified search result to increase the recall of the search results. Accordingly, if the input document is introduced via OCR or if the input document is known to be a partial document and the search results have provided a match with few if any additional documents (e.g., revised, or related), then the document determined to be a match is processed as the input document at to identify additional documents (i.e., to increase the recall of the original search results).

Kubota such not teach or suggest repeating a search using the matching document to identify similar documents if the second set of documents includes a matching document but no similar documents, as recited in the subject claims. The Examiner cites col. 3, lines 63-66 of Kubota to teach this limitation. (A comparison document stored in the storage medium "[can be], in the case of multiple documents...a set of documents including the input document, or a set of document [sic] extracted by search or the like. The contents of document [sic] may be of a natural language or a program language."). This citation is improper since storing the input document as a comparison document does not teach or suggest a document matching the input document that is employed to identify similar documents when no similar documents are originally located. This mechanism is in place to allow for a greater number of search results and is not contemplated or disclosed by Kubota.

For at least the aforementioned reasons, Kubota does not teach or suggest the subject invention as recited in independent claim 1 (or claims 10-13, and 15 which depend therefrom). Accordingly, withdrawal of this rejection is respectfully requested.

First Obviousness Rejection

The examiner has rejected claims 3-7 and 16-21 under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Gilfillan et al. (US PG Pub. No. 2002/0165856). This rejection should be withdrawn for at least the following reasons. Kubota in view of Gilfillan et al. do not teach or suggest the subject invention as set forth in the subject claims.

Independent claim 16 (and similarly independent claims 18 and 20) recites a method for computing ratings of keywords extracted from an input document. A determination is made as to whether each keyword in the list of keywords exists in a domain specific dictionary of words. Lists of keywords are identified for each output document in the first set of documents by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages. A measure of similarity is computed between the input document and each output document in the first set of documents. A second set of documents is defined with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value. The list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document. The query is repeated until a predetermined number of results are obtained or the query is terminated. If the second set of documents includes a matching document but no similar documents, the above steps are repeated using the matching document to identify similar documents. Kubota and Gilfillan, individually or in combination, do not teach or suggest such claimed aspects of the subject invention.

In particular, neither Kubota nor Gilfillan teach or suggest tokenizing one or more keywords employed in a search at one or more predefined word boundaries while maintaining order of the sequence of input text, as recited in the subject claims. As known, tokenization is the process of taking categorized blocks of text, usually consisting of indivisible characters known as lexemes that are categorized according to function, giving them meaning. Kubota does not teach such tokenization. Instead Kubota teaches extracting unique character strings from an input document and a performing a similarity search by using the unique character string. The extraction of the unique character string is performed by calculating and evaluating an amount of feature of a character string through comparison between appearance frequency appearing in the input document and appearance frequency in a set of documents to be searched. See, e.g., Abstract. There is no mention of taking categorized blocks of text according to function to give them meaning. Accordingly, Kubota does not contemplate tokenization of any keywords to perform this search.

Moreover, neither Kubota nor Gilfillan teach or suggest the translation of keywords into one or more languages, as recited in the subject claims. Rather, Kubota performs searches based on character strings from documents.. Such documents are presented in a single language and Kubota is silent regarding expanding a search to include translations of the selected character strings. Gilfillan does not make up for such deficiencies.

Additionally, Kubota does not teach or suggest repeating a query until a predetermined number of results are obtained or the query is terminated. Instead, Kubota teaches determining whether all fixed length chains created from the search character string have been searched. If so, the process proceeds to a next step. If not, the process returns to a previous step where the search process is performed for the next fixed length chain by using the character chain file. Thus, Kubota teaches performing searches until all search terms have been searched. There is no mechanism disclosed in either Kubota or Gilfillan that contemplates performing such a search until a predetermined number of results have been located or until a search is terminated.

In addition, Kubota and Gilfillan do not mention if the second set of documents includes a matching document but no similar documents repeating a search using the matching document to identify similar documents. The general measure of quality of search results has two components: precision and recall. Search results have good precision when the document being searched for is identified. Search results have good recall when not only the document being search for is identified but also all copies of the document being search for are identified. In this manner, in evaluating search results for input documents that were introduced via OCR or known to be a partial document, search results tend to have good precision but poor recall. When this occurs, as recited in the subject claims, the search is repeated using the matching document to identify similar documents when the second set of documents includes a matching document but no similar documents. The search is performed using the identified search result to increase the recall of the search results. Accordingly, if the input document is introduced via OCR or if the input document is known to be a partial document and the search results have provided a match with few if any additional documents (e.g., revised, or related), then the document determined to be a match is processed as the input document at to identify additional documents (i.e., to increase the recall of the original search results).

Neither Kubota nor Gilfillan teach or suggest repeating a search using the matching document to identify similar documents if the second set of documents includes a matching document but no similar documents, as recited in the subject claims. The Examiner cites col. 3, lines 63-66 of Kubota to teach this limitation. (A comparison document stored in the storage medium "[can be], in the case of multiple documents...a set of documents including the input document, or a set of document [sic] extracted by search or the like. The contents of document [sic] may be of a natural language or a program language."). This citation is improper since storing the input document as a comparison document does not teach or suggest a document matching the input document that is employed to identify similar documents when no similar documents are originally located. This mechanism is in place to allow for a greater number of search results and is not contemplated or disclosed by Kubota or Gilfillan.

For at least the aforementioned reasons, Kubota in combination with Gilfillan does not teach or suggest the subject invention as recited in independent claims 16, 18, or 20 (or claims 17, 19, and 21 which respectively depend therefrom). Moreover, claims 3-4 and 6-7 depend from independent claim 1 and Gilfillan does not make up for the aforementioned deficiencies of Kubota regarding identification of a second set of documents (e.g., output documents) as being one of a match, a revision, and a relation of the input document. Accordingly, withdrawal of this rejection is respectfully requested.

Second Obviousness Rejection

The examiner has rejected claims 8-9 under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Withgott et al. (US Patent No. 5,748,805). This rejection should be withdrawn for at least the following reasons. Claims 8-9 depend from independent claim 1, and Withgott et al. does not make up for the aforementioned deficiencies of Kubota regarding identification of lists of keywords for output documents in a first set of documents by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages or defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value wherein the query is repeated until a predetermined number of results are obtained or the query is terminated, if the second set of documents includes a matching document but no similar documents, the search is repeated using the matching document to identify similar documents. Thus, for at least the reasons discussed above with respect to claim 1, the combination of Kubota and Withgott et al. do not teach or suggest the subject claims. Accordingly, the rejection of this claim should be withdrawn.

Third Obviousness Rejection

The examiner has rejected claim 14 under 35 U.S.C. §103(a) as being unpatentable over Kubota in view of Cofino et al. (US PG Pub No. 2005/0187931). This rejection should be withdrawn for at least the following reasons. Claim 14 depends from independent claim 1, and Cofino et al. does not make up for the aforementioned deficiencies of Kubota regarding identification of lists of keywords for output documents in a first set of documents by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages or defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value wherein the query is repeated until a predetermined number of results are obtained or the query is terminated, if the second set of documents includes a matching document but no similar documents, the search is repeated using the matching document to identify similar documents. Thus, for at least the reasons discussed above with respect to claim 1, the combination of Kubota and Cofino et al. do not teach or suggest the subject claim. Accordingly, the rejection of these claims should be withdrawn.

CONCLUSION

For the reasons detailed above, it is submitted all claims remaining in the application (Claims 1, and 3-21) are now in condition for allowance. The foregoing comments do not require unnecessary additional search or examination.

No additional fee is believed to be required for this Amendment. However, the undersigned attorney of record hereby authorizes the charging of any necessary fees, other than the issue fee, to Xerox Deposit Account No. 24-0037.

In the event the Examiner considers personal contact advantageous to the disposition of this case, he/she is hereby authorized to call Mark Svat, at Telephone Number (216) 861-5582.

Respectfully submitted,

FAY SHARPE LLP

~~1/18/07~~ 3/19/07
Date

~~Mark Svat, Reg. No.~~
Kevin M. Dunn, Reg. No. 52,842
1100 Superior Avenue, Seventh Floor
Cleveland, OH 44114-2579
216-861-5582

Certificate of Mailing or Transmission

I hereby certify that this correspondence (and any item referred to herein as being attached or enclosed) is (are) being

☒ deposited with the United States Postal Service "Express Mail" service under 37 CFR 1.10, addressed to: Mail Stop RCE, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date indicated below.

□ transmitted to the USPTO by electronic transmission via EFS-Web on the date indicated below.

Express Mail Label No.: EV 889470985 US	Signature:
Date: ~~1/18/07~~ 3/19/07	Name: Kevin M. Dunn